FORM PTO-1390 (Modified) (REV 11-2000) U.S. DEPARTMENT OF COMMERCE PATENT AND TRADEMARK OFFICE

TRANSMITTAL LETTER TO THE UNITED STATES DESIGNATED/ELECTED OFFICE (DO/EO/US) CONCERNING A FILING UNDER 35 U.S.C. 371

ATTORNEY'S DOCKET NUMBER	U.S. APPLICATION NO. (IF KNOWN, SEE 37 CFR)
A32313-PCT USA	09/913745

INTERNATIONAL APPLICATION NO.	INTERNATIONAL FILING DATE	PRIORITY DATE CLAIMED
PCT/US00/04118	18 February 2000	19 February 1999

TITLE OF INVENTION

MULTI-DOCUMENT SUMMARIZATION SYSTEM AND METHOD

APPLICANT(S) FOR DO/EO/US

MCKEOWN, Kathleen R. and BARZILAY, Regina

Applicant herewith submits to the United States Designated/Elected Office (DO/EO/US) the following items and other information:

1. ☒ This is a **FIRST** submission of items concerning a filing under 35 U.S.C. 371.
2. ☐ This is a **SECOND** or **SUBSEQUENT** submission of items concerning a filing under 35 U.S.C. 371.
3. ☐ This is an express request to begin national examination procedures (35 U.S.C. 371(f)). The submission must include itens (5), (6), (9) and (24) indicated below.
4. ☒ The US has been elected by the expiration of 19 months from the priority date (Article 31).
5. ☒ A copy of the International Application as filed (35 U.S.C. 371 (c) (2))
   a. ☒ is attached hereto (required only if not communicated by the International Bureau).
   b. ☐ has been communicated by the International Bureau.
   c. ☐ is not required, as the application was filed in the United States Receiving Office (RO/US).
6. ☐ An English language translation of the International Application as filed (35 U.S.C. 371(c)(2)).
   a. ☐ is attached hereto.
   b. ☐ has been previously submitted under 35 U.S.C. 154(d)(4).
7. ☒ Amendments to the claims of the International Application under PCT Article 19 (35 U.S.C. 371 (c)(3))
   a. ☐ are attached hereto (required only if not communicated by the International Bureau).
   b. ☐ have been communicated by the International Bureau.
   c. ☐ have not been made; however, the time limit for making such amendments has NOT expired.
   d. ☒ have not been made and will not be made.
8. ☐ An English language translation of the amendments to the claims under PCT Article 19 (35 U.S.C. 371(c)(3)).
9. ☐ An oath or declaration of the inventor(s) (35 U.S.C. 371 (c)(4)).
10. ☐ An English language translation of the annexes of the International Preliminary Examination Report under PCT Article 36 (35 U.S.C. 371 (c)(5)).
11. ☒ A copy of the International Preliminary Examination Report (PCT/IPEA/409).
12. ☒ A copy of the International Search Report (PCT/ISA/210).

Items 13 to 20 below concern document(s) or information included:

13. ☐ An Information Disclosure Statement under 37 CFR 1.97 and 1.98.
14. ☐ An assignment document for recording. A separate cover sheet in compliance with 37 CFR 3.28 and 3.31 is included.
15. ☐ A **FIRST** preliminary amendment.
16. ☐ A **SECOND** or **SUBSEQUENT** preliminary amendment.
17. ☐ A substitute specification.
18. ☐ A change of power of attorney and/or address letter.
19. ☐ A computer-readable form of the sequence listing in accordance with PCT Rule 13ter.2 and 35 U.S.C. 1.821 - 1.825.
20. ☐ A second copy of the published international application under 35 U.S.C. 154(d)(4).
21. ☐ A second copy of the English language translation of the international application under 35 U.S.C. 154(d)(4).
22. ☒ Certificate of Mailing by Express Mail
23. ☒ Other items or information:

Form PCT/IB/301; Form PCT/IB/304; Form PCT/IPEA/401; Communication Requesting Change in Applicant's Nationality; Form PCT/IPEA/402; Form PCT/IB/308; Form PCT/IB/332; a postcard and a check in the amount of $363.

Express Mail Label No.: EK839853032US Date of Deposit: August 16, 2001

JC02 Rec'd PCT/PTO 16 AUG 2001

U.S. APPLICATION NO. (IF KNOWN, SEE 37 CFR	INTERNATIONAL APPLICATION NO.	ATTORNEY'S DOCKET NUMBER
09/913745	PCT/US00/04118	A32313-PCT USA

24. The following fees are submitted:.

BASIC NATIONAL FEE (37 CFR 1.492 (a) (1) - (5)) :

		CALCULATIONS PTO USE ONLY
☐ Neither international preliminary examination fee (37 CFR 1.482) nor international search fee (37 CFR 1.445(a)(2)) paid to USPTO and International Search Report not prepared by the EPO or JPO	$1000.00	
☐ International preliminary examination fee (37 CFR 1.482) not paid to USPTO but International Search Report prepared by the EPO or JPO	$860.00	
☐ International preliminary examination fee (37 CFR 1.482) not paid to USPTO but international search fee (37 CFR 1.445(a)(2)) paid to USPTO	$710.00	
☒ International preliminary examination fee (37 CFR 1.482) paid to USPTO but all claims did not satisfy provisions of PCT Article 33(1)-(4)	$690.00	
☐ International preliminary examination fee (37 CFR 1.482) paid to USPTO and all claims satisfied provisions of PCT Article 33(1)-(4)	$100.00	
ENTER APPROPRIATE BASIC FEE AMOUNT =		$690.00
Surcharge of **$130.00** for furnishing the oath or declaration later than ☐ 20 ☐ 30 months from the earliest claimed priority date (37 CFR 1.492 (e)).		$0.00

CLAIMS	NUMBER FILED	NUMBER EXTRA	RATE		
Total claims	22 - 20 =	2	x $18.00	$36.00	
Independent claims	3 - 3 =	0	x $80.00	$0.00	
Multiple Dependent Claims **(check if applicable)**.			☐	$0.00	
TOTAL OF ABOVE CALCULATIONS =				$726.00	
☒ Applicant claims small entity status. (See 37 CFR 1.27). The fees indicated above are reduced by 1/2.				$363.00	
SUBTOTAL =				$363.00	
Processing fee of **$130.00** for furnishing the English translation later than ☐ 20 ☐ 30 months from the earliest claimed priority date (37 CFR 1.492 (f)). +				$0.00	
TOTAL NATIONAL FEE =				$363.00	
Fee for recording the enclosed assignment (37 CFR 1.21(h)). The assignment must be accompanied by an appropriate cover sheet (37 CFR 3.28, 3.31) **(check if applicable).** ☐				$0.00	
TOTAL FEES ENCLOSED =				$363.00	
				Amount to be: refunded	$
				charged	$

a. ☒ A check in the amount of **$363.00** to cover the above fees is enclosed.

b. ☐ Please charge my Deposit Account No. _____ in the amount of _____ to cover the above fees. A duplicate copy of this sheet is enclosed.

c. ☒ The Commissioner is hereby authorized to charge any additional fees which may be required, or credit any overpayment to Deposit Account No. **02-4377** A duplicate copy of this sheet is enclosed.

d. ☐ Fees are to be charged to a credit card. **WARNING:** Information on this form may become public. **Credit card information should not be included on this form.** Provide credit card information and authorization on PTO-2038.

NOTE: Where an appropriate time limit under 37 CFR 1.494 or 1.495 has not been met, a petition to revive (37 CFR 1.137(a) or (b)) must be filed and granted to restore the application to pending status.

SEND ALL CORRESPONDENCE TO:

Henry Tang
BAKER BOTTS LLP
30 Rockefeller Plaza
New York, NY 10112-0228

SIGNATURE

Paul D. Ackerman
NAME

39,891
REGISTRATION NUMBER

August 16, 2001
DATE

A32313-PCT USA 070050.1589

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants	:	MCKEOWN, Kathleen R. et al.
Serial No.	:	To be assigned
Filed	:	18 February 2000
For	:	MULTI-DOCUMENT SUMMARIZATION SYSTEM AND METHOD

EXPRESS MAIL CERTIFICATION

Express Mail Mailing No. EK839853032US

Date of Deposit - August 16, 2001

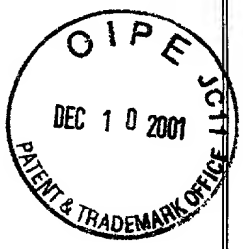
I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 C.F.R. §1.10 on the date indicated above and is addressed to: Box PCT, Assistant Commissioner for Patents, Washington, D.C., 20231.

Signature of person mailing correspondence

Name of person mailing correspondence: Leroy Chick

BAKER BOTTS L.L.P.

Attorney Docket Number: 32313-PCT-USA-070050.1589



Title:

MULTI-DOCUMENT SUMMARIZATOIN SYSTEM AND METHOD

Use Space Below for Additional Information:

MULTI-DOCUMENT SUMMARIZATION SYSTEM AND METHOD

SPECIFICATION

Statement of Government Rights

The United States Government may have certain rights to the invention set forth herein pursuant to a grant by the National Science Foundation, Contract No. IRI-96-18797.

Statement of Related Applications

This application claims the benefit of United States provisional patent application, Serial No. 60/120,659, entitled "Information Fusion in the Context of Multi-Document Summarization," which was filed on February 19, 1999.

Field of the Invention

The present invention relates generally to information summarization and more particularly relates to systems and methods for generating a summary for a set of multiple, related documents.

Background of the Invention

The amount of information available today drastically exceeds that of any time in history. With the continuing expansion of the Internet, this trend will likely continue well into the future. Often, people conducting research of a topic are faced with information overload as the number of potentially relevant documents exceeds the researchers ability to individually review each document. To address this problem, information summaries are often relied on by researchers to quickly evaluate a document to determine if it is truly relevant to the problem at hand.

Given the vast collection of documents available, there is interest in developing and improving the systems and methods used to summarize information content. For individual documents, domain-dependent template based systems and

domain-independent sentence extraction methods are known. Such known systems can provide a reasonable summary of a single document. However, these systems are not able to compare and contrast related documents in a document set to provide a summary of the collection.

The ability to summarize collections of documents containing related information is desirable to further expedite the research process. For example, for a researcher interested in news stories regarding a certain event, a summary of all documents from a given source, or multiple sources, would provide a valuable overview of the documents within the set. From such a summary, the researcher may be able to extract the information desired, or at the very least, make an informed decision regarding the relevance of the set of documents. Therefore, there remains a need for systems and methods which can generate a summary of related documents in a document set.

Summary of the Invention

It is an object of the present invention to provide a system and method for generating a summary of a set of multiple, related documents.

It is a further object of the present invention to provide a system and method for generating a summary of a set of multiple, related documents which use paraphrasing rules to detect similarities in non-identical phrases in the documents.

A present method for generating a summary of related documents in a collection includes extracting phrases from the documents which have common focus elements. Phrase intersection analysis is performed on the extracted phrases to generate a phrase intersection table. Temporal processing can be performed on the phrases in the phrase intersection table to remove ambiguous temporal references and to sort the phrases in a temporal sequence. Sentence generation is performed using the phrases in the phrase intersection table to generate the multidocument summary.

Preferably, the phrase intersection analysis operation can include representing the phrases in tree structures having root nodes and children nodes; selecting those tree structures with verb root nodes; comparing the selected root nodes to the other root nodes to identify identical nodes; applying paraphrasing rules to non-

identical root nodes to determine if non identical nodes are equivalent; and evaluating the children nodes of those tree structures where the parent nodes are identical or equivalent. The tree structure can take the form of a DSYNT tree structure. The paraphrasing rules can include one or more rules which are selected from the group consisting of ordering of sentence components, main clause versus a relative clause, different syntactic categories, change in grammatical features, omission of an empty head, transformation of one part of speech to another, and semantically related words.

In an embodiment of the present method, the temporal processing includes time stamping phrases based on a first occurrence of the phrase in the collection; substituting date certain references for ambiguous temporal references; ordering the phrases based on the time stamp; and inserting a temporal marker if a temporal gap between phrases exceeds a threshold value.

Preferably, a phrase divergence processing operation can also be performed to include phrases that signal changes in focus of the documents in the collection.

Sentence generation can includes mapping the phrases, represented in the tree structure, to an input format of a language generation engine and then operating the language generation engine.

A present system for generating a summary of a collection of related documents includes a storage device for storing the documents in the collection, a lexical database, such as WordNet, and a processing subsystem operatively coupled to the storage device and the lexical database. The processing subsystem is programmed to perform multiple document summarization including: accessing the documents in the storage device; using the lexical database to extract phrases from the documents with similar focus elements; performing phrase intersection analysis on the extracted phrases to generate a phrase intersection table; performing temporal processing on the phrases in the phrase intersection table; and performing sentence generation using the phrases in the phrase intersection table.

The methods described above can be encoded in the form of a computer program stored in computer readable media, such as CD-ROM, magnetic storage and the like.

Brief Description of the Drawing

Further objects, features and advantages of the invention will become apparent from the following detailed description taken in conjunction with the accompanying figures showing illustrative embodiments of the invention, in which

Figure 1 is a flow chart illustrating the operation of the present multiple document summarization system;

Figure 2 is a flow chart of a phrase intersection processing operation in accordance with the system operation of Figure 1;

Figure 3 is a pictorial diagram of a DSYNT tree structure for an exemplary sentence;

Figure 4 is a flow chart of a temporal processing operation in accordance with the system operation of Figure 1;

Figure 5 is a simplified block diagram of an embodiment of the present multiple document summarization system.

Throughout the figures, the same reference numerals and characters, unless otherwise stated, are used to denote like features, elements, components or portions of the illustrated embodiments. Moreover, while the subject invention will now be described in detail with reference to the figures, it is done so in connection with the illustrative embodiments. It is intended that changes and modifications can be made to the described embodiments without departing from the true scope and spirit of the subject invention as defined by the appended claims.

Detailed Description of Preferred Embodiments

Figure 1 is a flow chart which provides an overview of the operation of the present multiple document summarization system. Initially, a set of documents, in computer readable format and grouped by a common theme or domain, is presented to the summarization system. From the collection of documents, entities are identified and sentences are extracted from the documents which are relevant to the focus of the articles. Entities can be identified and extracted in a number of ways, such as by use of an information extraction engine. A suitable information extraction engine is TALENT, which is available from International Business Machines, Inc. In step 100,

phrases are extracted from the documents which include terms that are present in at least two of the documents. In addition, divergent phrases, which may be indicative of contrasts in the documents, are also extracted from the document in step 110. Following extraction, phrase intersection processing (step 120) and phrase divergence processing (step 130) are performed to evaluate and compare the extracted phrases and determine whether such phrases should be included in the resulting multiple document summary. Since phrases are extracted from multiple documents and can include temporal references which are ambiguous when taken out of context from the original document, temporal processing (step 140) is performed on the phrases selected for the summary. Finally, sentence generation (step 150) is used to transform selected phrases into a coherent summary.

Figure 2 is a flow chart which further illustrates steps that can be performed in connection with phrase intersection processing of step 120. The selected phrases are grammatically parsed and represented in a tree structure, such as a DSYNT tree diagram, which is generally known in the art. An example of such a diagram is illustrated in Figure 3. The parse trees can be generated by a conventional grammatical parser, such as Collin's parser. The DSYNT tree structure is a way of representing the constituent dependencies resulting from a predicate-argument sentence structure. In the tree structure, each non-auxiliary word in the sentence has a node which is connected to its direct dependents. Grammatical features of each word are also stored in the node. To facilitate subsequent comparisons, words in the nodes are kept in their canonical form.

Returning to Figure 2, those trees which have root nodes which are verbs are selected and used as the basis for comparison. Each such verb based tree is compared against the other trees derived from the sentences extracted from the documents in the collection (step 220). A comparison is made to determine if two nodes are identical (step 230). If two nodes are identical, those nodes are added to an output tree (step 235) and the nodes are evaluated to determine if there are further nodes descending from the root node (step 240). Such further nodes are referred to as children nodes. If children nodes are present (step 245), the comparison in step 230 is repeated for each of child node. If the analysis of the children nodes is complete at

step 240, a determination is made as to whether the trees with common root nodes represent a phrase intersection (step 250). For example, if there is commonality in the root node and at least two children nodes of that root node, that phrase can be deemed complete and added to a phrase intersection table (step 255). If no phrase intersection is detected at step 250, the next parent node is selected for processing (step 260) and control returns to step 230.

Returning to step 230, if two nodes are not identical, it is still possible for the nodes to be equivalent. To make this determination, the present method employs a set of paraphrasing rules to evaluate the nodes (step 265). Paraphrasing, which can be broadly defined as alternative ways a human speaker can choose to "say the same thing" by using linguistic knowledge, generally occurs at a "surface" level, e.g., it is achieved by using semantically related words and syntactic transformations.

In the case of a set of related documents, theme sentences of the documents will generally be close semantically. This limits the scope of different paraphrasing types to be evaluated. From an analysis of paraphrasing patterns evaluated through themes of a training corpus derived from TDT, the following non-exhaustive set of paraphrasing categories have been found to occur with the greatest frequency:

1. ordering of sentence components: "*Tuesday they met...*" and "*They met ... Tuesday*";
2. main clause vs. a relative clause: "*...a building was devastated by the bomb*" and "*...a building, devastated by the bomb*";
3. realization in different syntactic categories, e.g., classifier vs. apposition: "*Palestinian leader Arafat*" and "*Arafat, Palestinian leader*", *"Pentagon speaker" and "speaker from the Pentagon*";
4. change in grammatical features: active/passive, time, number. "*...a building was devastated by the bomb*" and "*...the bomb devastated a building*";
5. omission of an empty head: "*group of students*" and "*students*";
6. transformation from one part of speech to another: "*building devastation*" and "*...building was devastated*"; and

7. using semantically related words such as synonyms: "*return*" and "*alight*", "*regime*" and "*government*".

The categories presented are used as paraphrasing rules by the present methods. The majority of these categories,such as ordering, can be fully implemented in an automatic way, . However, some of the rules can be only approximated to a certain degree in an automated system. For example, identification of similarity based on semantic relations between words depends on the scope of coverage of the thesaurus employed. Word similarity can be established using relationships such as synonymy, hyponymy/hypernymy, and meronymy/holonymy which are detectable using the WordNet language database which is described in the article "WordNet: A lexical Database for English", by G.A. Miller, Communications of the ACM, Vol. 38, No. 11. pp. 39-41, November 1995.

If any of the included paraphrasing rules are satisfied for non-identical nodes, the nodes are deemed equivalent (step 270). Equivalent nodes are added to the output tree (step 235) and processed in the same manner as identical nodes. If no paraphrasing rule is applicable to non-identical nodes, there is no phrase intersection with the current tree (step 280).

In addition to phase intersection processing, which compares phrases for similarity, it is also desirable to perform phrase divergence processing (step 130), which compares selected phrases for differences. Phrase divergence may indicate a critical change in the course of events through a set of related documents and would be worthy of inclusion in a summary. For example, a collection of articles regarding a plane crash could begin with a focus on the passengers as "survivors" and later refer to "casualties," "victims," "bodies" and the like, which signify a turning point in the events described by the documents. WordNet can also be used in phrase divergence processing by evaluating focus relationships such antonymy (e.g., "happiness is opposite to sadness").

Once phrases are selected from the documents for the summary, temporal processing can be performed to sequence the phrases and eliminate ambiguous temporal references. The flow chart of Figure 4 illustrates an overview of the temporal processing operations performed in the present methods. Using a rule

that an event is assumed to have occurred on the day that it is first reported, a time stamp can be applied to the selected phrases based on the earliest occurrence of the phrase in the collection of documents (step 405). In certain cases, phrases may include ambiguous temporal references, such as today, yesterday, etc. In this case, such ambiguous references can be replaced by a date certain reference, such as by changing "Yesterday it was reported...." to "*On 01/02/2000*, it was reported...". Such substitutions, which are performed in step 410, can be implemented using the Emacs "calendar" package.

The extracted phrases can then be ordered in accordance with the assigned date stamp (step 415). In certain cases, a large temporal gap may exist between consecutive phrases. In such a case, if the gap exceeds a threshold, such as two days, a temporal marker can be inserted between the phrases to indicate this gap in time (step 420). This may be significant, for example, in the case of a collection of news articles where the gap in time can also correspond to a change in focus in the articles.

With the phrases selected and sorted in temporal order, sentence generation (step 150) can be performed to synthesize a coherent summary. Sentence generation involves two major operations. First, the DSYNT representation of the phrases to be used in sentence generation are mapped to the appropriate syntax of a selected language generation engine. Then, the language generation engine is operated to arrange the phrases into coherent sentences. A suitable language generation engine is FUF/SURGE, which is available from Columbia University, New York , New York, as well as from Ben Gurion University, Department of Computer Science, Beer-Sheva, Israel. The acronym FUF stands for Functional Unification Formalism interpreter and the acronym SURGE stands for syntactic realization grammar for text generation. The input specification for the FUF/SURGE engine includes a semantic role, *circumstantial,* which itself includes a temporal feature. The inclusion of the semantic attributes enables FUF/SURGE to perform various paraphrasing operations to the input phrases to improve the resulting sentences.

Figure 5 is a simplified block diagram of a multiple document summarization system in accordance with the present invention. The system 500

includes a processor section 505 wherein the processing operations set forth in Figure 1 are performed. The system also includes non-volatile storage coupled to the processor section 505 for document storage 510, collection summary storage 515, lexical database storage 518 and program storage 520. Generally these storage systems are read/write data storage systems, such as magnetic media and read/write optical storage media. However, the document collection storage may take the form of read-only storage, such as a CD-ROM storage device. The system further includes RAM memory 525 coupled to the processor section for temporary storage during operation. The system 500 will generally include one or more input device 530 such as a keyboard, digitizer, mouse and the like, which is coupled to the processor section 505. Similarly, a conventional display device 535 is generally provided which is also operatively coupled to the processor section.

The particular hardware embodiment is not critical to the practice of the present invention. Various computer platforms and architectures can be used to implement the system 500, such as personal computers, workstations, networked computers, and the like. The functions described in the system can be performed locally or in a distributed manner, such as over a local area network or the Internet. For example, the document collection storage 510 may be at a remote archive location which is accessed by the processor section 505 via a connection to the Internet.

Although the present invention has been described in connection with specific exemplary embodiments, it should be understood that various changes, substitutions and alterations can be made to the disclosed embodiments without departing from the spirit and scope of the invention as set forth in the appended claims.

CLAIMS

1. A method for generating a summary of a plurality of related documents in a collection comprising:

extracting phrases having focus elements from the plurality of documents;

performing phrase intersection analysis on the extracted phrases to generate a phrase intersection table;

performing temporal processing on the phrases in the phrase intersection table; and

performing sentence generation using the phrases in the phrase intersection table.

2. The method of generating a summary as defined by claim 1, wherein the phrase intersection analysis comprises:

representing the phrases in tree structures having root nodes and children nodes;

selecting those tree structures with verb root nodes;

comparing the selected root nodes to the other root nodes to identify identical nodes;

applying paraphrasing rules to non-identical root nodes to determine if non identical nodes are equivalent; and

evaluating the children nodes of those tree structures where the parent nodes are identical or equivalent.

3. The method of claim 2, wherein the tree structure is a DSYNT tree structure.

4. The method of claim 2, wherein the paraphrasing rules are selected from the group consisting of ordering of sentence components, main clause versus a relative clause, different syntactic categories, change in grammatical features, omission of an

empty head, transformation of one part of speech to another, and semantically related words.

5. The method of claim 1, wherein the temporal processing includes:

time stamping phrases based on a first occurrence of the phrase in the collection;

substituting date certain references for ambiguous temporal references;

ordering the phrases based on the time stamp; and

inserting a temporal marker if a temporal gap between phrases exceeds a threshold value.

6. The method of claim 1, further comprising a phrase divergence processing operation.

7. The method of claim 1, wherein the sentence generation includes mapping phrases to an input format of a language generation engine and operating the language generation engine.

8. A system for generating a summary of a plurality of related documents in a collection comprising:

a storage device for storing the documents in the collection;

a lexical database; and

a processing subsystem, the processing subsystem being operatively coupled to the storage device and the lexical database, the processing subsystem being programmed to access the documents in the storage device and:

using the lexical database to extract phrases having focus elements from the plurality of documents;

performing phrase intersection analysis on the extracted phrases to generate a phrase intersection table;

performing temporal processing on the phrases in the phrase intersection table; and

performing sentence generation using the phrases in the phrase intersection table.

9. The system for generating a summary as defined by claim 9, wherein the phrase intersection analysis processing further comprises:

representing the phrases as data structures having root nodes and children nodes;

selecting those data structures with verb root nodes;

comparing the selected root nodes to the other root nodes to identify identical nodes;

applying paraphrasing rules to non-identical root nodes to determine if non identical nodes are equivalent; and

evaluating the children nodes of those tree structures where the parent nodes are identical or equivalent.

10. The system of claim 9, wherein the data structure is a DSYNT tree structure.

11. The system of claim 9, wherein the paraphrasing rules are selected from the group consisting of ordering of sentence components, main clause versus a relative clause, different syntactic categories, change in grammatical features, omission of an empty head, transformation of one part of speech to another, and semantically related words.

12. The system of claim 8, wherein the temporal processing includes:

time stamping phrases based on a first occurrence of the phrase in the collection;

substituting date certain references for ambiguous temporal references;

ordering the phrases based on the time stamp; and

inserting a temporal marker if a temporal gap between phrases exceeds a threshold value.

13. The system of claim 8, further comprising a phrase divergence processing operation.

14. The system of claim 8, wherein the processing subsystem includes a language generation engine and wherein sentence generation includes mapping phrases to an input format of the language generation engine and then operating the language generation engine.

15. The system of claim 8, wherein the storage device for storing the documents in the collection is remotely located from the processing subsystem.

16. A computer readable media for programming a computer system to perform a method of generating a summary of a plurality of related documents in a collection comprising:

extracting phrases having focus elements from the plurality of documents;

performing phrase intersection analysis on the extracted phrases to generate a phrase intersection table;

performing temporal processing on the phrases in the phrase intersection table; and

performing sentence generation using the phrases in the phrase intersection table.

17. The computer readable media of claim 16, wherein the phrase intersection analysis comprises:

representing the phrases in tree structures having root nodes and children nodes;

selecting those tree structures with verb root nodes;

comparing the selected root nodes to the other root nodes to identify identical nodes;

applying paraphrasing rules to non-identical root nodes to determine if non identical nodes are equivalent; and

evaluating the children nodes of those tree structures where the parent nodes are identical or equivalent.

18. The computer readable media of claim 17, wherein the tree structure is a DSYNT tree structure.

19. The computer readable media of claim 17, wherein the paraphrasing rules are selected from the group consisting of ordering of sentence components, main clause versus a relative clause, different syntactic categories, change in grammatical features, omission of an empty head, transformation of one part of speech to another, and semantically related words.

20. The computer readable media of claim 16, wherein the temporal processing includes:

time stamping phrases based on a first occurrence of the phrase in the collection;

substituting date certain references for ambiguous temporal references;

ordering the phrases based on the time stamp; and

inserting a temporal marker if a temporal gap between phrases exceeds a threshold value.

21. The computer readable media of claim 16, further comprising a phrase divergence processing operation.

22. The computer readable media of claim 16, wherein the sentence generation includes mapping phrases to an input format of a language generation engine and operating the language generation engine.

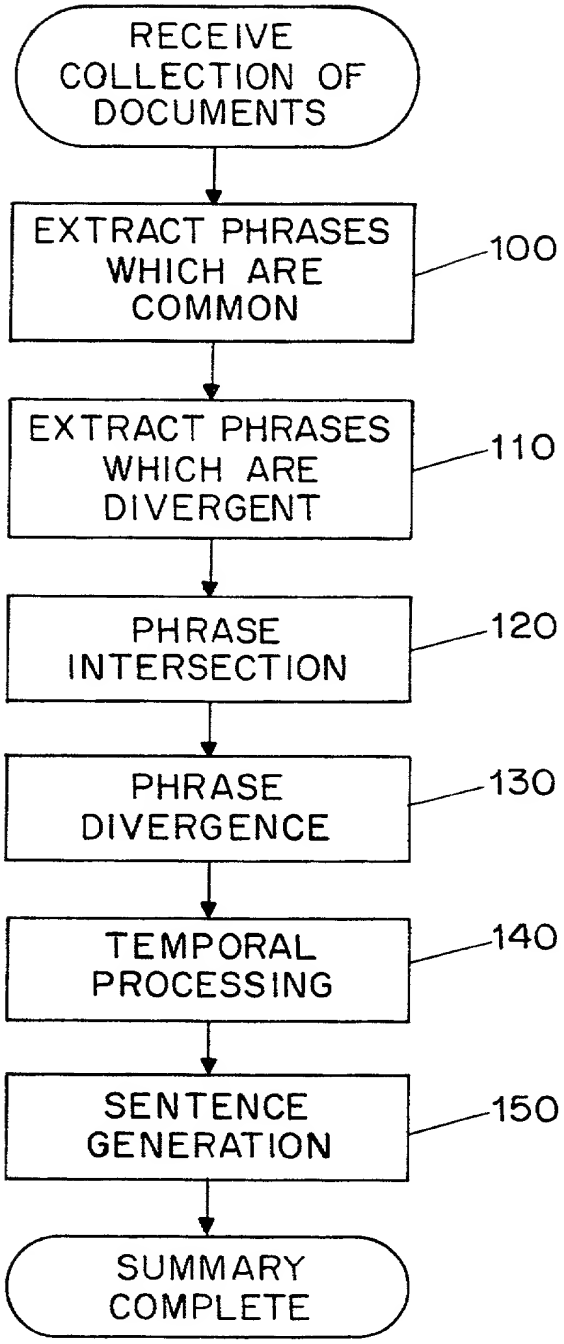

RECEIVE COLLECTION OF DOCUMENTS
EXTRACT PHRASES WHICH ARE COMMON
100
EXTRACT PHRASES WHICH ARE DIVERGENT
110
PHRASE INTERSECTION
120
PHRASE DIVERGENCE
130
TEMPORAL PROCESSING
140
SENTENCE GENERATION
150
SUMMARY COMPLETE

FIG. 1

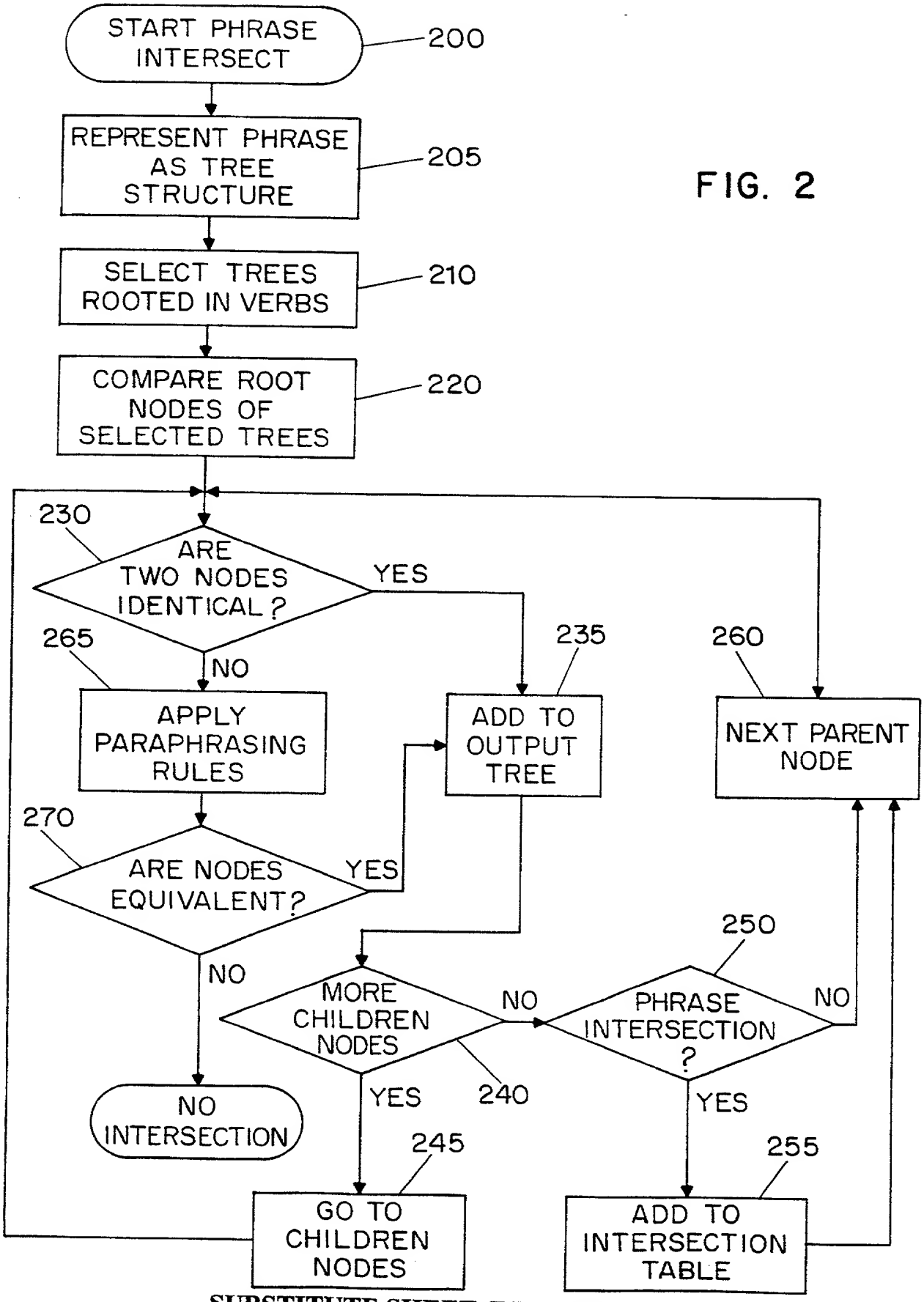

START PHRASE INTERSECT
200
REPRESENT PHRASE AS TREE STRUCTURE
205
SELECT TREES ROOTED IN VERBS
210
COMPARE ROOT NODES OF SELECTED TREES
220
230
ARE TWO NODES IDENTICAL?
YES
NO
265
APPLY PARAPHRASING RULES
235
ADD TO OUTPUT TREE
260
NEXT PARENT NODE
270
ARE NODES EQUIVALENT?
YES
NO
250
MORE CHILDREN NODES
NO
PHRASE INTERSECTION ?
NO
YES
240
NO INTERSECTION
YES
245
GO TO CHILDREN NODES
255
ADD TO INTERSECTION TABLE

FIG. 2

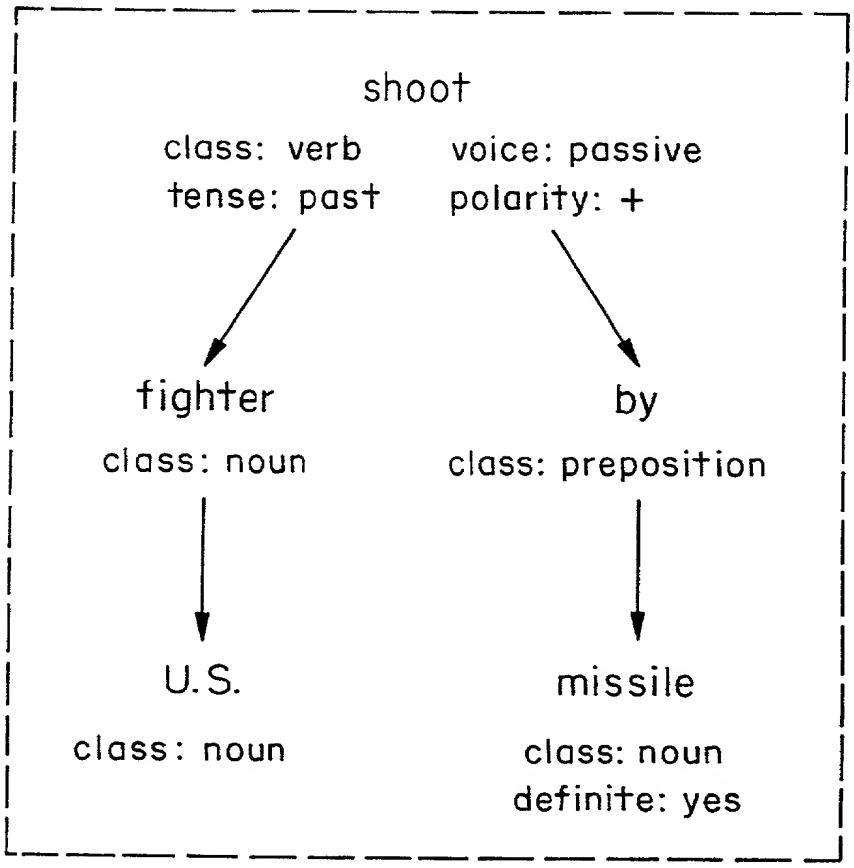

shoot
class: verb
tense: past
voice: passive
polarity: +
fighter
class: noun
by
class: preposition
U.S.
class: noun
missile
class: noun
definite: yes

FIG. 3

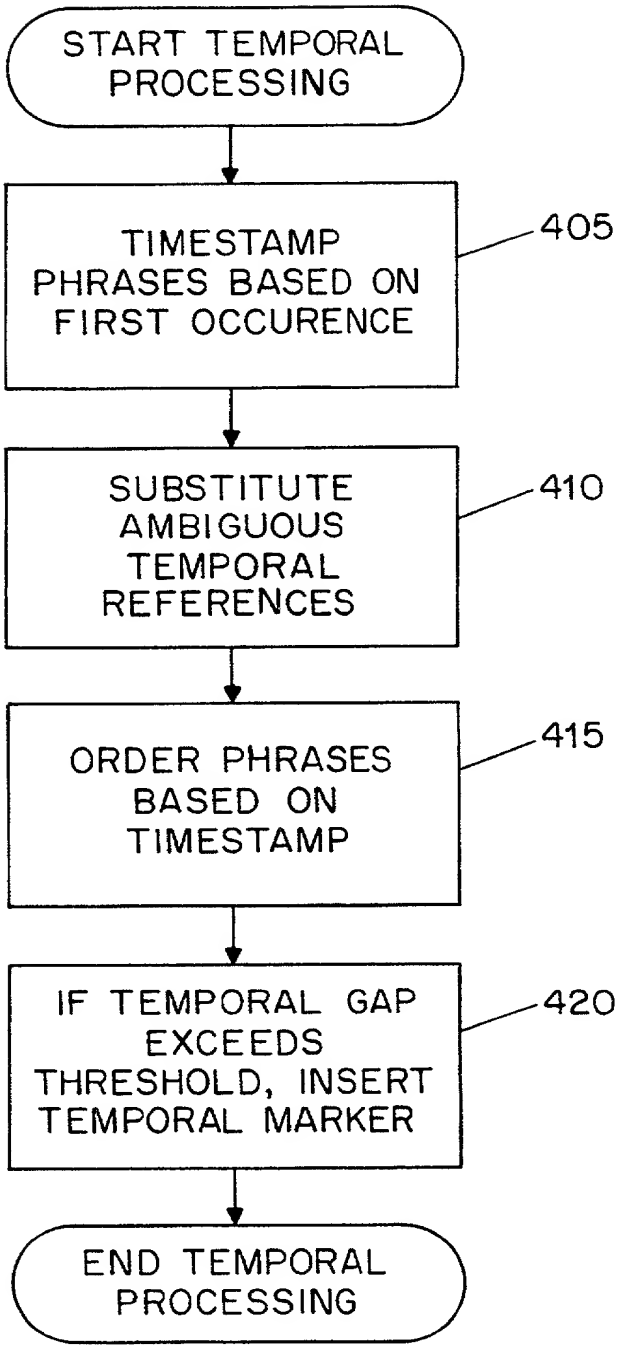

START TEMPORAL PROCESSING
TIMESTAMP PHRASES BASED ON FIRST OCCURENCE
405
SUBSTITUTE AMBIGUOUS TEMPORAL REFERENCES
410
ORDER PHRASES BASED ON TIMESTAMP
415
IF TEMPORAL GAP EXCEEDS THRESHOLD, INSERT TEMPORAL MARKER
420
END TEMPORAL PROCESSING

FIG. 4

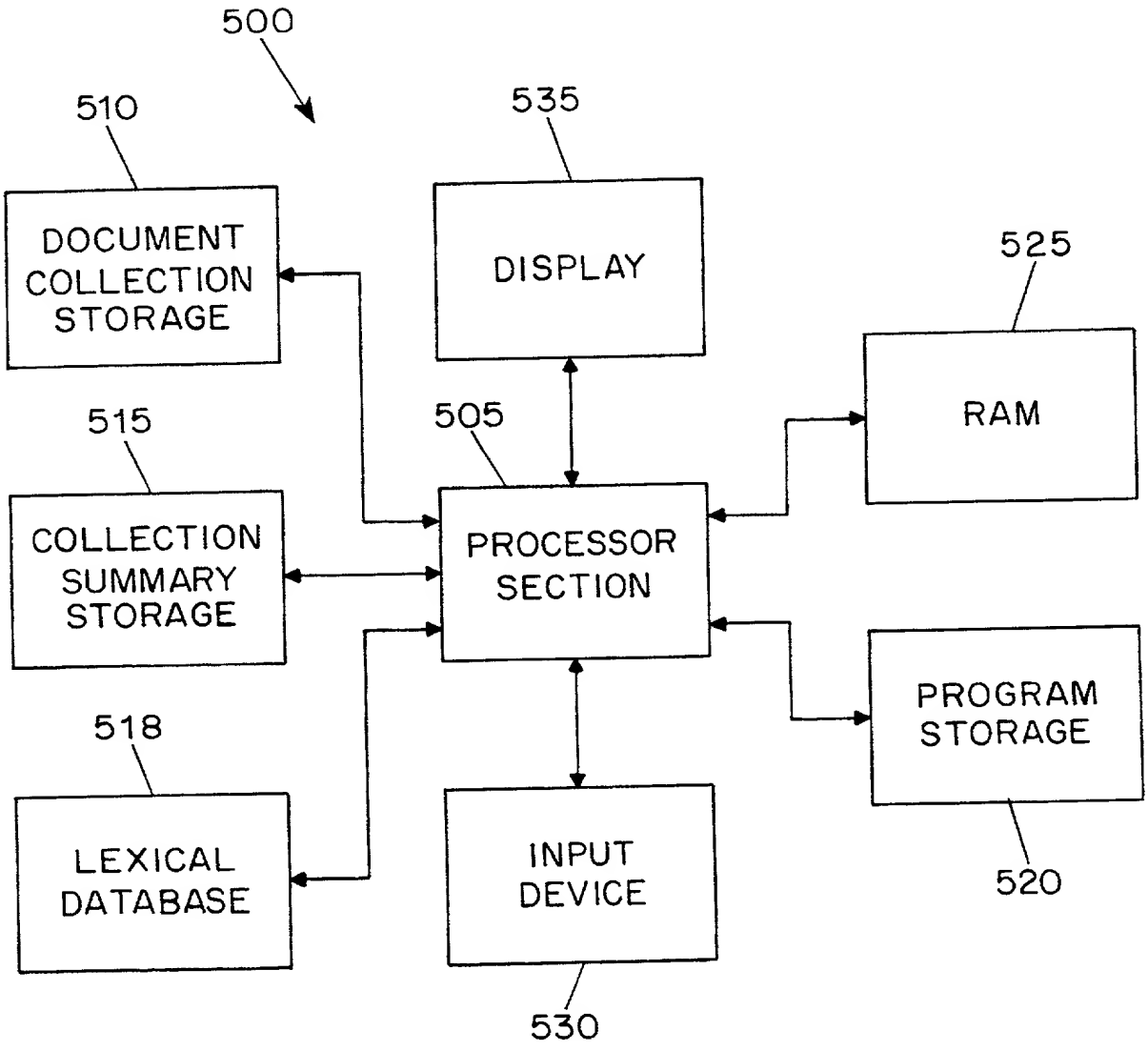
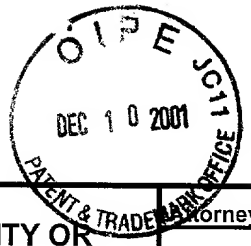

500
510
DOCUMENT COLLECTION STORAGE
535
DISPLAY
525
RAM
515
COLLECTION SUMMARY STORAGE
505
PROCESSOR SECTION
518
LEXICAL DATABASE
INPUT DEVICE
PROGRAM STORAGE
520
530

FIG. 5

#4

BAKER BOTTS L.L.P.



DECLARATION FOR UTILITY OR DESIGN PATENT APPLICATION (37 CFR 1.63)

☐ Declaration Submitted with Initial Filing **OR** ☑ Declaration Submitted after Initial Filing (surcharge (37 CFR 1.16 (e)) required)

Attorney Docket Number	32313-PCT-USA-070050.1589
First Named Inventor	KATHLEEN R. MCKEOWN
COMPLETE IF KNOWN	
Application Number	09/913,745
Filing Date	August 16, 2001
Group Art Unit	
Examiner Name	

As a below named inventor, I hereby declare that:

My residence, mailing address, and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

MULTI-DOCUMENT SUMMARIZATION SYSTEM AND METHOD

(Title of the Invention)

the specification of which

☐ is attached hereto

OR

☑ was filed on (MM/DD/YYYY) 08/16/2001 as United States Application Number or PCT International Application Number 09/913,745 and was amended on (MM/DD/YYYY) _____ (if applicable).

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment specifically referred to above.

I acknowledge the duty to disclose information which is material to patentability as defined in 37 CFR 1.56, including for continuation-in-part applications, material information which became available between the filing date of the prior application and the national or PCT international filing date of the continuation-in-part application.

I hereby claim foreign priority benefits under 35 U.S.C. 119(a)-(d) or (f), or 365(b) of any foreign application(s) for patent, inventor's or plant breeder's rights certificate(s), or 365(a) of any PCT international application which designated at least one country other than the United States of America, listed below and have also identified below, by checking the box, any foreign application for patent, inventor's or plant breeder's rights certificate(s), or any PCT international application having a filing date before that of the application on which priority is claimed.

Prior Foreign Application Number(s)	Country	Foreign Filing Date (MM/DD/YYYY)	Priority Not Claimed	Certified Copy Attached? YES	NO
			☐	☐	☐
			☐	☐	☐
			☐	☐	☐
			☐	☐	☐

☐ Additional foreign application numbers are listed on a supplemental priority data sheet PTO/SB/02B attached hereto:

DECLARATION Utility or Design Patent Application

Claim for Benefit of Prior U.S. Provisional Application(s)

I hereby claim the benefit under Title 35, United States Code, § 119(e) of any United States provisional application(s) listed below:

Provisional Application Number	Filing Date

Claim for Benefit of Earlier U.S./PCT Application(s) under 35 U.S.C. 120

(complete this part only if this is a divisional, continuation or C-I-P application)

I hereby claim the benefit under Title 35, United States Code, § 120 of any United States application(s) or PCT international application(s) designating the United States of America that is/are listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior application(s) in the manner provided by the first paragraph of Title 35, United States Code § 112, I acknowledge the duty to disclose information as defined in Title 37, Code of Federal Regulations, Section 1.56 which occurred between the filing date of the prior applications(s) and the national or PCT international filing date of this application:

Application Number	Filing Date	Status (patented, pending, abandoned)
PCT/US00/04118	February 19, 1999	Pending

BAKER BOTTS LLP

Attorney Docket Number 32313-PCT-USA-070050.1589

DECLARATION Utility or Design Patent Application

Direct all correspondence to: ☑ Customer Number or Bar Code Label 21003 OR ☐ Correspondence address below

Name

Address

City	State	ZIP
Country	Telephone	Fax

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 18 U.S.C. 1001 and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

NAME OF SOLE OR FIRST INVENTOR : ☐ A petition has been filed for this unsigned inventor

Given Name (first and middle [if any])	Family Name or Surname
KATHLEEN R.	MCKEOWN

Inventor's Signature Kathleen R. McKeown Date 11/16/01

Residence: City	State	Country	Citizenship
WAYNE	NEW JERSEY	UNITED STATES	UNITED STATES NJ

Mailing Address 20 PROSPECT ROAD

City	State	ZIP	Country
WAYNE	NEW JERSEY	07470	UNITED STATES

NAME OF SECOND INVENTOR: ☐ A petition has been filed for this unsigned inventor

Given Name (first and middle [if any])	Family Name or Surname
REGINA	BARZILAY

Inventor's Signature Date 11/19

Residence: City	State	Country	Citizenship
NEW YORK	NEW YORK	UNITED STATES	UNITED STATES NY

Mailing Address 548 RIVERSIDE DRIVE, APT. 4B

City	State	ZIP	Country
NEW YORK	NEW YORK	10027	UNITED STATES

☐ Additional inventors are being named on the ____supplemental Additional Inventor(s) sheet(s) PTO/SB/02A attached hereto.

#4

BAKER BOTTS L.L.P.

Please type a plus sign (+) inside this box → +

POWER OF ATTORNEY OR AUTHORIZATION OF AGENT

Application Number	09/913,745
Filing Date	August 16, 2001
First Named Inventor	KATHLEEN R. MCKEOWN
Group Art Unit	
Examiner Name	
Attorney Docket Number	32313-PCT-USA-070050.1589

I hereby appoint:

☑ Practitioners at Customer Number 21003 →

Place Customer Number Bar Code Label here

OR

☐ Practitioner(s) named below:

Name	Registration Number

as my/our attorney(s) or agent(s) to prosecute the application identified above, and to transact all business in the United States Patent and Trademark Office connected therewith.

Please change the correspondence address for the above-identified application to:

☑ The above-mentioned Customer Number.

OR

☐ Firm *or* Individual Name					
Address					
Address					
City		State		Zip	
Country					
Telephone		Fax			

I am the:

☑ Applicant/Inventor.

☐ Assignee of record of the entire interest. See 37 CFR 3.71.
Statement under 37 CFR 3.73(b) is enclosed. (Form PTO/SB/96).

SIGNATURE of Applicant or Assignee of Record

Name	KATHLEEN R. MCKEOWN
Signature	Kathleen R. McKeown
Date	11/16/01

NOTE: Signatures of all the inventors or assignees of record of the entire interest or their representative(s) are required. Submit multiple forms if more than one signature is required, see below*.

☐ *Total of _____ forms are submitted.

BAKER BOTTS LLP

Please type a plus sign (+) inside this box → ✚

POWER OF ATTORNEY OR AUTHORIZATION OF AGENT

Application Number	09/913,745
Filing Date	August 16, 2001
First Named Inventor	KATHLEEN R. MCKEOWN
Group Art Unit	
Examiner Name	
Attorney Docket Number	32313-PCT-USA-070050.1589

I hereby appoint:

☑ Practitioners at Customer Number 21003 → *Place Customer Number Bar Code Label here*

OR

☐ Practitioner(s) named below:

Name	Registration Number

as my/our attorney(s) or agent(s) to prosecute the application identified above, and to transact all business in the United States Patent and Trademark Office connected therewith.

Please change the correspondence address for the above-identified application to:

☑ The above-mentioned Customer Number.

OR

☐ Firm *or* Individual Name					
Address					
Address					
City		State		Zip	
Country					
Telephone		Fax			

I am the:

☑ Applicant/Inventor.

☐ Assignee of record of the entire interest. See 37 CFR 3.71.
Statement under 37 CFR 3.73(b) is enclosed. (Form PTO/SB/96).

SIGNATURE of Applicant or Assignee of Record

Name	REGINA BARZILAY
Signature	
Date	11/19

NOTE: Signatures of all the inventors or assignees of record of the entire interest or their representative(s) are required. Submit multiple forms if more than one signature is required, see below*.

☐ *Total of _____ forms are submitted.